J·T·L·A

The Journal of Technology, Learning, and Assessment

Volume 2, Number 6 · February 2004

Developing Computerized Versions of Paper-and-Pencil Tests: Mode Effects for Passage-Based Tests

Mary Pommerich

www.jtla.org

A publication of the Technology and Assessment Study Collaborative
Caroline A. & Peter S. Lynch School of Education, Boston College

J·T·L·A

Volume 2, Number 6

Developing Computerized Versions of Paper-and-Pencil Tests: Mode Effects for Passage-Based Tests

Mary Pommerich

Editor: Michael Russell
russelmh@bc.edu
Technology and Assessment Study Collaborative
Lynch School of Education, Boston College
Chestnut Hill, MA 02467

Copy Editor: Kathleen O'Connor
Design and Layout: Thomas Hoffmann

JTLA is a free on-line journal, published by the Technology and Assessment Study Collaborative, Caroline A. & Peter S. Lynch School of Education, Boston College.

Preferred citation:

Pommerich, M. (2004). Developing computerized versions of paper-and-pencil tests: Mode effects for passage-based tests. *Journal of Technology, Learning, and Assessment, 2*(6). Available from http://www.jtla.org

Abstract:

As testing moves from paper-and-pencil administration toward computerized administration, how to present tests on a computer screen becomes an important concern. Of particular concern are tests that contain necessary information that cannot be displayed on screen all at once for an item. Ideally, the method of presentation should not interfere with examinee performance on the test. Examinees should perform similarly on an item regardless of the mode of administration. This paper discusses the development of a computer interface for passage-based, multiple-choice tests. Findings are presented from two studies that compared performance across computer and paper administrations of several fixed-form tests. The effect of computer interface changes made between the two studies is discussed. The results of both studies showed some performance differences across modes. Evaluations of individual items suggested a variety of factors that could have contributed to mode effects. Although the observed mode effects were in general small, overall the findings suggest that it would be beneficial to develop an understanding of factors that can influence examinee behavior and to design a computer interface accordingly, to ensure that examinees are responding to test content rather than features inherent in presenting the test on computer.

J·T·L·A

Developing Computerized Versions of Paper-and-Pencil Tests: Mode Effects for Passage-Based Tests

Introduction

As testing moves from paper-and-pencil administration toward computerized administration, how to transfer tests from a test booklet to a computer screen becomes an important concern. Computerized administration is perhaps less of an issue for tests with discrete items in which individual items can be presented in full on a computer screen. Transferring tests of this nature from booklet to computer may be a relatively straight-forward process. Computerized administration is likely more of an issue for passage-based tests with content that can be viewed in full on a two-page spread in a booklet, but cannot be presented on a single computer screen.

Developing a computer interface for a passage-based test is a complicated process. As such, it is important to develop an understanding of the presentation choices we make and how they can affect an examinee's performance, particularly if computer and paper versions of a test will co-exist. In a dual-platform testing program with tests that cannot easily be transferred to computer, taking certain items in one mode or the other could possibly advantage some examinees. Even in a computer-only platform, decisions about how to present the test could affect examinee performance. Seemingly subtle differences in how the test is presented on computer could have a not-so-subtle effect on examinee performance.

Wang and Kolen (2001) argue that for practical reasons, it is often desirable to maintain score comparability across paper and computer adaptive administrations, even though doing so may result in the loss of some of the potential advantages of a computer adaptive test. Because of potential mode effects, Parshall, Spray, Kalohn, and Davey (2002) suggest that testing programs that treat scores across different administration platforms as equivalent should perform studies to document the comparability of the test scores. Since the advent of computerized testing, a multitude of comparability studies have been conducted on a variety of types of tests, typically to compare scores across computer and paper administrations of the same fixed-form multiple-choice tests. Results have been mixed across the studies.

The research does generally seem to indicate, however, that the more complicated it is to present or take the test on computer, the greater the possibility of

mode effects. For tests where all of the information for an item could be presented in its entirety on screen, results of comparability studies often showed small or insignificant mode effects (Bridgeman, Lennon, & Jackenthal, 2003; Choi & Tinkler, 2002; Hetter, Segall, & Bloxom, 1997; Bergstrom, 1992; Spray, Ackerman, Reckase, & Carlson, 1989). For tests where all of the information for an item could not be presented in its entirety on screen, and some form of navigation (typically scrolling) was necessary on the part of computer examinees to view all of the information, results often showed more significant mode effects (Bridgeman, Lennon, & Jackenthal, 2003; Choi & Tinkler, 2002; Bergstrom, 1992). Reading tests with passages requiring navigation typically tended to show more mode effects than mathematics tests. These findings imply that tests that require navigation may be more subject to mode effects than tests that do not require navigation. For tests that required a written response from examinees (i.e., an open-ended assessment or a performance writing assessment), substantial mode effects were noted (Russell, 1999; Russell & Haney, 1997). Altogether, the results of previous comparability studies seem to suggest that mode differences might not be tied so much to test content per se, but rather, to the degree to which the presentation of the test and the process of taking the test differ across the modes of administration.

Because computer technology is continually changing, testing programs should conduct their own comparability studies using their own tests and technology, as comparability results might not generalize beyond a given test and computer interface. Likewise, it is important for testing programs to conduct their own comparability studies, as results do not always turn out as might be expected. For example, Mazzeo, Druesne, Raffeld, Checketts, and Muhlstein (1991) compared performance of what they considered to be relatively simple to present multiple-choice items across computer and paper administrations of mathematics and English composition tests. Most items were presented in their entirety on the computer screen, yet mode differences were found at the test-score level. After modifying the computer interface in response to the findings, mode effects were still noted on the mathematics test, but not for the English composition test.

In evaluating mode effects, it is useful to look not only at comparability at the total score level, but also at the item level, because there can be strong mode effects for individual items that cancel out at the overall score level. Most comparability studies have not examined mode effects at the item level. One exception is found in Schaeffer, Reese, Steffen, McKinley, and Mills (1993), who concluded there were no substantive mode effects across individual items on the GRE tests. Item level evaluations can also provide insights into sources of mode differences and can help develop an understanding of how examinees interact with item features when presented in a test booklet versus a computer interface. Muter (1996) recommends taking into account individual differences in designing interfaces by evaluating user differences, isolating the source of variation, and redesigning the interface to accommodate differences among users. Mazzeo et al. (1991) demonstrated the use of an iterative process for developing a computer interface by making changes to the interface in response to initial comparability findings and then evaluating comparability again. Because the modified interface still resulted in mode effects,

the authors noted that they would continue to modify the computer interface in an attempt to eliminate mode effects.

This paper discusses the development of a computer interface for passage-based multiple-choice tests that require examinees to navigate through the passage while responding to an item. The interface development was done in two stages in the hopes of developing an understanding of the effect that interface features can have on computer examinees' performance and how computer examinees perform relative to paper examinees. An initial interface was developed and a comparability study was conducted. Modifications were made to the interface in response to the study findings and evaluations of interface and booklet features, item characteristics, and examinee behaviors for some individual items. Another comparability study was then conducted using the same test forms.

Performances are compared across paper and computer modes and across interface variations, for each comparability study. Results are summarized at the total score level and for individual items. Some factors that might have contributed to mode differences or affected test performance in general are discussed. The effects of interface changes made between the two studies are also discussed. Two specific questions are addressed:

1. Do examinees respond to items in the same way across administration modes and computer interface variations on passage-based tests that require navigation?
2. Where mode effects are observed, what are some of the factors contributing to the mode effects?

Comparability Studies: Tests and Interfaces

Comparability studies were performed in 1998 and 2000, Comparability 1 and Comparability 2, respectively. Participants in each study were in Grades 11 and 12. In each study, the same fixed-form tests were administered across paper-and-pencil and computer modes in the content areas of English, Reading, Science Reasoning, and Mathematics. In addition, identical forms were administered in English, Reading, and Science Reasoning across the two studies. Different sets of items were included in the Mathematics test across the two comparability studies. As such, results for Mathematics are not presented.[1] An initial computer interface was used in Comparability 1 and then modified for Comparability 2 based on findings from Comparability 1. The interface used in Comparability 1 is referred to as Interface 1. The interface used in Comparability 2 is referred to as Interface 2.

English Test

Test Content

The English test consisted of four passages containing underlined words and phrases, with 15 multiple-choice items per passage (60 items total). For most items, examinees were instructed to choose the response option for the underlined portion that best expressed the idea, made the statement appropriate for standard written English, or was worded most consistently with the style and tone of the passage as a whole. These types of items had no stimulus associated with them (i.e., there were only response options, and no preceding question). For some items, there was a stimulus present that asked a question about the underlined portion in the passage, and examinees were instructed to choose the best answer to the question.

Booklet Presentation

In the booklet presentation of the English test, the passage and items were presented jointly on a page. The passage was presented on the left half of the page, while the items were presented on the right half of the page. Each underlined portion was always aligned with the top of the corresponding item (i.e., with the stimulus, or with the first response option for items with no stimulus). Each passage and accompanying items occupied about two booklet pages. Examinees were able to move freely throughout all English passages and items in the booklet while taking the English test. They could respond to items and passages in any order, and were not required to give responses to all items.

Computer Presentation

In the computer presentation for both Interface 1 and Interface 2, a 17-inch monitor was used, set to a resolution of 1280 × 1024. The passage and an individual item were presented jointly on the screen, with the passage appearing in a window on the left half and the item appearing in a window on the right half of the screen. The passage was not visible in its entirety on the computer screen. The examinee had to scroll to see the complete passage, although the passage automatically scrolled for examinees on various items (discussed further below). Examinees moved between items by clicking on a specific item number or by using "Next Question" or "Previous Question" buttons. Within a passage, examinees were allowed to answer items in any order. They were required to answer all items prior to moving on to the next passage. Once an examinee completed a passage and moved on to the next passage, they were not allowed to return to the previous passage. Also, passages were presented one at a time, so that examinees could not see the next passage until they proceeded to it. A similar presentation of the passage and item windows was used with the computerized Reading and Science Reasoning tests, along with the same rules for moving between items and passages.

Comparability 1 Interface Features

In Interface 1, the following features were utilized:

- The full underlined portion in the passage window was highlighted using a yellow color.
- The passage automatically scrolled when an item was selected for which the corresponding underlined portion in the passage was not visible on screen (about every 6th item). Examinees could also scroll line-by-line or manipulate a sliding scroll bar to move quickly through the passage, although further scrolling beyond the automatic scrolling typically was not necessary.
- The underlined portions were not aligned with the top of the item (in contrast to the booklet presentation).

Effect of Comparability 1 Results on Comparability 2 Interface

Results for the English test in Comparability 1 (to be discussed in more detail later) showed that the test as a whole tended to favor computer examinees, although some individual items did favor paper examinees. After a review of the test content, test booklet, computer interface, and interviews with examinees, the following hypotheses were posited as possible explanations for why computer examinees performed better overall than paper examinees (see also Pommerich and Burden (2000) for further discussion):

- The use of color highlighting for the full underlined portion was advantageous to computer examinees as it provided an additional cue to the underlined portion.
- On some items, computer examinees were better able to focus on relevant sections of passages/items because those sections were centered in the passage and item windows and examinees were not distracted by extraneous information presented in the rest of the test. This phenomenon will be referred to as the "focus effect."

Results also suggested the following hypotheses:

- Computer examinees might have been less likely to read the stimulus preceding the response options, for items containing a stimulus.
- Where the underlined portion was aligned with the response options might influence the response selected.

Thus, the following changes were implemented in Interface 2 for the English test:

- The color highlighting of the full underlined portion was removed. Instead, only the item number underneath the underlined portion was highlighted with color.
- The item number was placed adjacent to the top of the item within the item window to match how the item number is presented in the booklets. (In Interface 1, the items were numbered outside of the item window.)

In addition, two different automatic scrolling variations were utilized in Interface 2:

1. The passage scrolled when an item was selected that was not visible on screen (about every 6th item), so that the underlined portion was not always aligned with the top of the item. This condition will be referred to as *English Semi*. This scrolling style was used in both Interface 1 and Interface 2.
2. The passage scrolled every time a new item was selected, so that the underlined portion was always aligned with the top of the item. This condition will be referred to as *English Auto*. This scrolling style was only used in Interface 2.

Reading Test

Test Content and Presentation

The Reading test consisted of four passages with 10 multiple-choice items per passage (40 items total). Examinees were instructed to read the passage and choose the best answer for each item. Items on the Reading test generally fell into two types: questions that required a global understanding of the passage and questions that required knowledge of specific information given in the passage. For global understanding questions, examinees typically had to make an inference from what they had read to answer the question. Some of the items had line references associated with them (i.e., the item stimulus contained the number of a line or lines in the passage which examinees were directed to read). In the booklet presentation, the reading passage was presented first in its entirety, in two columns per page. The passages were followed by the test items. Each passage and accompanying items occupied about two booklet pages. Examinees were allowed to respond to items and passages in any order and could move freely throughout the paper-and-pencil version of the Reading test. The computer presentation for Reading corresponded to that described for the English test.

Comparabilty 1 Interface Features

In Interface 1, the following features were utilized:

- Examinees moved through the passage by scrolling.
- Examinees could scroll line-by-line, or manipulate a sliding scroll bar to move quickly through the passage.

Other relevant characteristics of the interface were as follows:

- Line-by-line scrolling speed was not very fast.
- Pre-test training for scrolling options was for line-by-line scrolling only. Examinees were not explicitly shown how to use the sliding scroll bar.
- Line breaks were not the same as in the booklet, so the content of referenced lines was not exactly the same across modes.

Effect of Comparability 1 Results on Comparability 2 Interface

Results for the Reading test in Comparability 1 (to be discussed in more detail later) showed that the test as a whole tended to favor paper examinees. After a review of the test content, test booklet, computer interface, and interviews with examinees, the following hypotheses were posited as possible explanations for why paper examinees performed better overall than computer examinees (see also Pommerich and Burden (2000) for further discussion):

- Paper examinees might have been more likely than computer examinees to experience "positional memory," whereby they remembered the location of information given in the passage, because the passage occurred in a fixed position on the page. Dillon (1992) suggests that readers of text presented on paper develop a visual memory for the location of information within the text, based on its spatial location both on the page and within the document.
- Computer examinees sometimes had difficulty locating information in the passage (recall that scrolling was the navigation method). Scrolling has been denigrated by some researchers as a means of navigating when reading continuous text on screen, because it allows only relative spatial orientation (Schwarz, Beldie, & Pastoor, 1983).
- Slow scrolling speed was an additional hindrance for computer examinees.
- Different line breaks across the paper and computer presentations could have created mode differences on questions with line references.

Thus, the following changes were implemented in Interface 2 for the Reading test:

- Line breaks for the passages were made the same across booklet and computer presentations, so that each line contained the same content across modes.
- Scrolling speed was increased.
- Prior to testing, examinees were explicitly taught to use the sliding scroll bar.

In addition, two different navigation variations were utilized in Interface 2:

1. Examinees moved through the passage by scrolling, using either line-by-line scrolling or a sliding scroll bar. This condition will be referred to as *Read Scroll*. Note that this scrolling was also used in Interface 1, although scrolling speed was slower and pre-test instruction on scrolling was less comprehensive than for Interface 2.
2. Examinees moved through the passage by paging. In this variation, the passage was divided into separate pages and the examinee moved between pages by clicking on a specific page number, or by using "Next Page" or "Previous Page" buttons. This condition will be referred to as *Read Page*. Paging was only used in Interface 2.

Science Reasoning Test

Test Content and Presentation

The Science Reasoning test consisted of seven passages with varying numbers of multiple-choice items per passage (5–7 items per passage; 40 items total). All of the passages contained figures and/or tables. In the booklet presentation, the passage was presented first in its entirety, followed by the test items. Each passage and accompanying items occupied about two booklet pages. Examinees taking the paper-and-pencil version could move freely throughout the passages and items in the booklet. The computer presentation for Science Reasoning corresponded to that described for the English test, with the additional feature that some figures and tables within the passage were enlargeable and moveable.

Comparability 1 Interface Features

In Interface 1, the following features were utilized:

- Examinees moved through the passage by scrolling.
- Examinees could scroll line-by-line, or manipulate a sliding scroll bar to move quickly through the passage.
- Some graphics (i.e., tables and figures) were enlargeable and moveable, and multiple graphics could be enlarged simultaneously and moved.

As described for Reading, Interface 1 for the Science Reasoning test was also characterized by:

- Line-by-line scrolling speed that was not very fast.
- Pre-test training for line-by-line scrolling only (examinees were not explicitly shown how to use the sliding scroll bar).

Effect of Comparability 1 Results on Comparability 2 Interface

Results for the Science Reasoning test in Comparability 1 (to be discussed in more detail later) showed some individual items favoring computer examinees and some individual items favoring paper examinees. Overall, there was no clear trend in results, although Passage 4 favored computer examinees, and the last passage (Passage 7) favored paper examinees. After a review of the test content, test booklet, computer interface, and interviews with examinees, the following hypotheses were posited as possible explanations for why computer and paper examinees performed differently on individual items/passages (see also Pommerich and Burden (2000) for further discussion):

- Paper examinees might have been more likely than computer examinees to experience "positional memory," whereby they remembered the location of information in the passage, because the passage occurred in a fixed position on the page.
- Computer examinees sometimes had difficulty locating information given in the passage (recall scrolling was the navigation method).
- Slow scrolling speed was an additional hindrance for computer examinees.
- Computer examinees had difficulty comparing information across tables or figures that did not appear on screen simultaneously (because of slow scrolling speed), and they were unaware that they could enlarge and move graphics so that graphics could be viewed simultaneously.
- Computer examinees were advantaged by a "focus effect" on some items (i.e., they were better able to focus on relevant sections of passages/items because those sections were centered in the passage and item windows and examinees were not distracted by extraneous information presented in the rest of the test).

Thus, the following changes were implemented in Interface 2 for the Science Reasoning test:

- Scrolling speed was increased.
- Prior to testing, examinees were explicitly taught to use the sliding scroll bar.

- Prior to testing, more explicit instructions were given on enlarging and moving graphics.

In addition, two different navigation variations were utilized in Interface 2:

1. Examinees moved through the passage by scrolling, using either a line-by-line scrolling or a sliding scroll bar. This condition will be referred to as *Science Scroll*. Note that this scrolling was also used in Interface 1, although scrolling speed was slower and pre-test instruction on scrolling was less comprehensive than for Interface 2.
2. Examinees moved through the passage by paging. In this variation, the passage was divided into separate pages and the examinee moved between pages by clicking on a specific page number, or by using "Next Page" or "Previous Page" buttons. This condition will be referred to as *Science Page*. Paging was only used in Interface 2.

Changes Between Interface 1 and Interface 2 for All Tests

The following changes were implemented between Interface 1 and Interface 2 the same way for each subject area test:

- The wording was changed on some buttons and on text adjacent to the buttons to be more concise and clear.
- Different colors and button designs were used to change the appearance of the interface.
- Additional passage and item numbering was added outside the passage and item windows, to clarify which item and passage the examinee was on (e.g., indicated Passage 1 of 4, Question 1 of 60). In Interface 1, the number of passages and total number of items were specified in the test directions. Once the test started, however, the current passage and item numbers were given, but no information was given as to how many passages or items remained.
- On startup of a passage, the first item was not displayed until the examinee selected the first item, to encourage examinees to read the passage before answering the first item.

Test Administration

Comparability 1: Participants and Test Forms

Comparability 1 compared performance across computer and paper-and-pencil administrations of the same fixed form, using computer Interface 1. Testing was conducted between September and December 1998. A total of 40 schools participated in the study, with approximately 8,600 eleventh and twelfth grade students tested overall. Within a school, examinees were randomly assigned to a paper-and-pencil or computer administration of a fixed-form test. Within each

administration mode, examinees were randomly assigned to one of the following content areas: English, Reading, Science Reasoning, or Mathematics. (Note that only one computer interface variation was used in each content area.) Thus, there were a total of eight administration conditions. All computer examinees took a short tutorial prior to testing that demonstrated how to use all of the functions necessary to take the computerized test (with the exception of demonstrating the use of the sliding scroll bar, as discussed earlier). The fixed-form tests used in the study were drawn from intact paper-and-pencil forms that had previously been administered operationally. The Reading and Science Reasoning forms were administered in their entirety with the same time constraints as used operationally, while a representative subset of items was selected from the English and Mathematics tests to accommodate a 35-minute testing period. Total testing time was 35 minutes for all content areas and modes.

Comparability 2: Participants and Test Forms

Comparability 2 compared performance across computer and paper-and-pencil administrations of the same fixed form, using computer Interface 2. Testing was conducted between October 2000 and January 2001. A total of 61 schools participated in the study, with approximately 12,000 eleventh and twelfth grade students tested. Within a school, examinees were randomly assigned to a paper-and-pencil or computer administration of a fixed-form test. Proportionately more examinees were assigned to the computer administration because of the interface variations. Examinees assigned to the paper mode were randomly assigned to one of the following content areas: English, Reading, Science Reasoning, or Mathematics. Examinees assigned to the computer mode were randomly assigned to one of the following content area and interface variations: English Auto, English Semi, Reading Scroll, Reading Page, Science Reasoning Scroll, Science Reasoning Page, or Mathematics. Thus, there were a total of 11 administration conditions.

All computer examinees took a short tutorial prior to testing that demonstrated how to use all of the functions necessary to take the computerized test. Different tutorials were used across the two comparability studies. The tutorial used in Comparability 2 was more comprehensive and more interactive than the tutorial used in Comparability 1. The same fixed-form tests were administered across computer and paper-and-pencil administration modes. With the exception of Mathematics, the forms used in Comparability 2 were identical to those used in Comparability 1. The Mathematics test was modified from Comparability 1 to include some new item types. Again, total test time for all content areas and modes was 35 minutes.

Results

Data Cleaning

Due to irregularities during assignment to a testing condition or during testing itself, some records were unusable; records that were problematic were deleted from the final analyses. The final sample sizes for the analyses are reported in Table 1.[2] The 8 groups within Comparability 1 are considered to be randomly equivalent, and the 11 groups within Comparability 2 are considered to be randomly equivalent.

Table 1 Final Sample Sizes for Analyses

	Comparability 1		Comparability 2	
Test	**Condition**	**N**	**Condition**	**N**
English	-	-	Computer Auto	1110
	Computer Semi	905	Computer Semi	1031
	Paper	1040	Paper	1137
Mathematics	Computer	918	Computer	1083
	Paper	994	Paper	1099
Reading	-	-	Computer Page	996
	Computer Scroll	908	Computer Scroll	1089
	Paper	985	Paper	1086
Science Reasoning	-	-	Computer Page	902
	Computer Scroll	827	Computer Scroll	1067
	Paper	947	Paper	1055

Completion Rates

A genuine concern in computerizing a paper-and-pencil test is that it might take more time for examinees to complete the test on computer than on paper. If it takes computer examinees longer than paper examinees to complete the same test, use of the same testing time across modes is potentially unfair to computer examinees. Many factors could contribute to an increased testing time for computer examinees. It may take more time to use a mouse to navigate and to respond to questions. It may take more time to find information on the computer if navigation is required. It may be more difficult to read from a computer screen than a test booklet. Research has suggested that reading speed (for both normal reading and skimming) is slower on a computer than in printed text (Muter, 1996; Muter & Maurutto, 1991). Reading comprehension, however, does not appear to be negatively affected by computer presentation (Sawaki, 2001; Dillon, 1992). Ideally, the computer interface would be designed so that it does not contribute to an increase in testing time on computer. However, if it does take more time to take the same

test on computer than on paper, then longer testing times may need to be allocated for testing on computer than testing on paper.

The percentages of examinees finishing the English, Reading, and Science Reasoning tests are given in Table 2, for both Comparability 1 and Comparability 2. For all content areas, the completion rates for the paper mode were lower in Comparability 2 than in Comparability 1. If there were no sample differences, we would expect completion rates for the paper mode to be about the same across studies because the same forms and testing time were used. The lower completion rates in Comparability 2 are likely a result of having a less academically able sample than in Comparability 1, arising from a greater solicitation of less academically able schools for Comparability 2.[3] We would expect then, that if our interface and tutorial changes did not have any effect on reducing the time needed to complete the test, completion rates for computer examinees in Comparability 2 would similarly be lower than completion rates for computer examinees in Comparability 1 simply because of the sample differences across the two studies.

Because the computer completion rates in Comparability 2 are about the same or higher than in Comparability 1, this suggests that the interface and tutorial changes in general decreased the amount of time needed to complete the test on computer. When comparing completion rates across computer and paper modes, it is important to note that completion rates might be inflated somewhat if examinees answer quickly and randomly at the very end of the test without reading the items, just to have a response to all items. It is likely easier for paper examinees to complete the test in such a way than computer examinees. As such, completion rates for paper examinees might appear higher than completion rates for computer examinees.

Table 2 Percent Completing the Test Across Comparability 1 and 2

Test	Comparability 1		Comparability 2	
	Condition	Percent Completing Test	Condition	Percent Completing Test
English	-	-	Computer Auto	83.42
	Computer Semi	81.00	Computer Semi	81.77
	Paper	81.00	Paper	78.36
Reading	-	-	Computer Page	64.16
	Computer Scroll	64.20	Computer Scroll	62.72
	Paper	76.20	Paper	70.99
Science Reasoning	-	-	Computer Page	65.19
	Computer Scroll	56.00	Computer Scroll	63.64
	Paper	68.80	Paper	60.86

English Test Completion Rates

Completion rates were the same for paper and computer examinees in Comparability 1. Completion rates for both computer conditions in Comparability 2 were higher than the paper completion rates. Completion rates were slightly higher for the Auto condition than for the Semi condition. There was a lot of white space in the English passage between adjacent lines, and because of the automatic scrolling examinees generally did not have to scroll while responding to individual items on computer. As a result, it might have been somewhat easier and quicker to focus on information on the computer than on paper if it was contained within the screen, and extraneous information was hidden from view. If the Comparability 2 completion rates were adjusted to account for the sample differences across the two studies (i.e., if the Comparability 2 paper and computer completion rates were each increased by the amount that would yield the same completion rates for paper across the two studies), then computer completion rates would be higher for Comparability 2 than for Comparability 1. The most probable explanation for this finding is that examinees in Comparability 2 had a greater awareness of where they were in the test and how much time remained than examinees in Comparability 1.

Reading Test Completion Rates

Completion rates were much higher for paper than for computer in Comparability 1. Completion rates were still higher for paper than for both computer conditions in Comparability 2, although the difference in completion rates was smaller, particularly for the Page condition. If the Comparability 2 completion rates were adjusted to account for the sample differences across the two studies, the Reading computer completion rates would be higher in Comparability 2 than they were in Comparability 1. The most probable explanations for this finding are that examinees in Comparability 2 had a greater awareness of where they were in the test and how much time remained than examinees in Comparability 1, and that navigation was improved in each interface variation.

The higher completion rates for paper than computer in Comparability 2 likely still occurred because the Reading passages were very dense, ranging from 723 to 873 words per passage. Each passage contained a lot of text that examinees had to review to find information, which may have been difficult to do on computer. It might be difficult to obtain similar rates of completion across paper and computer administrations for this test, without making some adjustment such as making the passages shorter, increasing testing time, or creating more white space between the lines of the passage. Adding more white space around the text would make it easier to read the passages on screen, but that would result in physically longer passages that would require more navigation, which could offset the advantage gained by adding white space, particularly for the scrolling variation. One possible solution was raised by Muter (1996), who suggests increasing spacing between lines while proportionately decreasing horizontal spacing between letters, to improve the clarity of text without affecting the length.

Science Reasoning Test Completion Rates

Completion rates were much higher for the paper condition than for the computer condition in Comparability 1, whereas in Comparability 2, the completion rates were higher for both of the computer conditions than for the paper condition. Completion rates were slightly higher for the Page condition than the Scroll condition. If the Comparability 2 completion rates were adjusted to account for sample differences across the two studies, computer condition completion rates would be substantially higher in Comparability 2 than Comparability 1. Again, the most probable explanations for this finding are a greater awareness for Comparability 2 examinees about where they were in the test and how much time remained, and improved navigation.

Improved navigation and greater awareness of the interface features are also plausible explanations for the higher completion rates for computer examinees over paper examinees in Comparability 2. Science Reasoning examinees sometimes had to compare information across figures or tables that were not visible simultaneously on the computer screen without enlarging or moving graphics. It is probable that many examinees did not utilize the enlarge feature and instead navigated back and forth between the graphics. Comparability 1 examinees that did not use the sliding scroll bar to move back and forth between tables and figures likely were severely hampered by the slow speed of the line-by-line scrolling under Interface 1. The improved navigation and improved instruction on using the navigation features of Interface 2 appears to have had a substantial effect on completion rates for Comparability 2 computer examinees.

The higher completion rates for Comparability 2 computer examinees relative to paper examinees might also be attributable in part to the "focus effect" posited earlier. It might be easier to focus on information on the computer than on paper if it is all contained within the screen, and extraneous information is hidden from view. The fact that the Science Reasoning results are the opposite from those observed for Reading (i.e., completion rates for Reading computer examinees in Comparability 2 are still below the completion rates of the paper examinees) suggests that even with improved navigation and training, there might be no focus effect for Reading computer examinees. This might be due to the nature of the information contained in the passage and how it is presented. While the passages are physically lengthy in Science Reasoning, the amount of text is much less than that of the Reading passages, and the inclusion of figures and tables creates more white space in the passage. Even if navigation were further improved, completion rates for Reading might never approach the completion rates for Science Reasoning because the sheer density of text contained in the Reading passage might make it more difficult to locate information.

Total Score Performance

Average total scores (and standard deviations) for each test and condition are given in Table 3, for both Comparability 1 and Comparability 2. The average scores were lower in Comparability 2 than in Comparability 1, as might be expected if the Comparability 2 examinees were less academically able. Table 4 gives the difference in average total scores across modes for each computer condition (computer – paper), and the value of the t-statistic for a test of the hypothesis that the average scores are equal across paper and computer modes. Positive values indicate a higher average score on computer than on paper.

Table 3 Average (and Standard Deviation) of Total Scores Across Comparability 1 and 2

	Comparability 1		Comparability 2	
Test	**Condition**	**Average (SD)**	**Condition**	**Average (SD)**
English	-	-	Computer Auto	34.03 (11.06)
60 Items	Computer Semi	36.09 (11.07)	Computer Semi	33.80 (11.06)
	Paper	34.90 (11.45)	Paper	32.38 (11.40)
Reading	-	-	Computer Page	20.16 (7.13)
40 Items	Computer Scroll	21.08 (7.17)	Computer Scroll	20.12 (6.91)
	Paper	22.13 (7.33)	Paper	20.37 (7.11)
Science Reasoning	-	-	Computer Page	21.97 (6.64)
	Computer Scroll	23.01 (6.91)	Computer Scroll	21.68 (6.80)
40 Items	Paper	23.07 (7.06)	Paper	21.24 (6.83)

Table 4 Difference in Average Total Scores Across Modes (Computer – Paper), and t-Statistic for Comparison of the Average Scores

	Comparability 1			Comparability 2		
Test	**Condition**	**Difference**	***t***	**Condition**	**Difference**	***t***
English	-	-	-	Auto	+1.65	+3.47**
	Semi	+1.19	+2.33*	Semi	+1.42	+2.94**
Reading	-	-	-	Page	-0.21	-0.66
	Scroll	-1.05	-3.14**	Scroll	-0.25	-0.82
Science Reasoning	-	-	-	Page	+0.73	+2.41*
	Scroll	-0.06	-0.17	Scroll	+0.44	+1.50

* $p < .05$
** $p < .01$

For English, computer examinees scored higher on average than paper examinees in Comparability 1, and the t-test showed a significant difference in average scores. In Comparability 2, computer examinees scored higher on average than paper examinees under both computer conditions, and the t-tests again showed significant differences in average scores for both conditions. Of the two computer conditions, Auto examinees scored slightly higher than Semi examinees. The difference in average scores across modes was larger for Comparability 2 than Comparability 1, so there was a widening of the performance gap favoring computer examinees across the two studies.

For Reading, computer examinees scored lower on average than paper examinees in Comparability 1, and the t-test showed a significant difference in average scores. In Comparability 2, computer examinees scored lower on average than paper examinees for both computer conditions. The t-tests did not show a significant difference in average scores for either condition. Of the two computer conditions, Page examinees scored slightly higher than Scroll examinees. The difference in average scores across modes was much smaller for Comparability 2 than Comparability 1, so there was a narrowing of the performance gap favoring paper examinees across the two studies.

For Science Reasoning, computer examinees scored slightly lower on average than paper examinees in Comparability 1. The t-test showed no significant difference in average scores. In Comparability 2, computer examinees scored higher on average than paper examinees for both computer conditions. Of the two computer conditions, Page examinees scored slightly higher than Scroll examinees. The t-test for the Page condition showed a significant difference in average scores; the t-test for the Scroll condition showed no significant difference in average scores. The difference in average scores across modes was larger for Comparability 2 than Comparability 1, with a shift in direction from slightly favoring paper examinees in Comparability 1 to favoring computer examinees in Comparability 2. This trend complements the shift in completion rates for Science Reasoning noted earlier.

For Reading and Science Reasoning, it is probable that slow navigation speed and lack of knowledge of navigation capabilities hindered Comparability 1 computer examinees' test performance. For all content areas, it is likely that improvements in Interface 2 and the tutorial made it easier for examinees to use the interface, navigate throughout the test, and respond more quickly, leading to an improved performance of computer examinees relative to paper examinees in Comparability 2 over Comparability 1. Although some of the t-test results were significant, it should be noted that no adjustments were made for multiple comparisons. Further, the effect sizes for the mean differences[4] observed across both studies were no larger than ± 0.15 in any content area, which is considered small by Cohen's (1988) standard.

Item Level Performance

Summaries of item level performance for English are shown in Figures 1–2. Each figure shows plots of individual item p-value differences (with error bands) across paper and computer conditions. Each passage is separated by a vertical line in the plots. Figure 1 shows the computer – paper p-value differences ± 2 standard errors for each computer condition from Comparability 2 (Semi and Auto). Figure 2 shows the computer – paper p-value differences ± 2 standard errors for Comparability 1 and for a baseline comparison based on two mutually exclusive random samples of examinees who took the items used in the comparability studies in a paper administration as part of a separate equating study. The two groups in the baseline comparison are considered to be randomly equivalent, so one group was arbitrarily assigned to represent the "computer" condition, while the other was assigned to represent the "paper" condition. The sample sizes for the computer and paper sample were fixed at those observed for the respective condition in Comparability 1.

(Figures 1 and 2 are shown on the following page.)

Figure 1

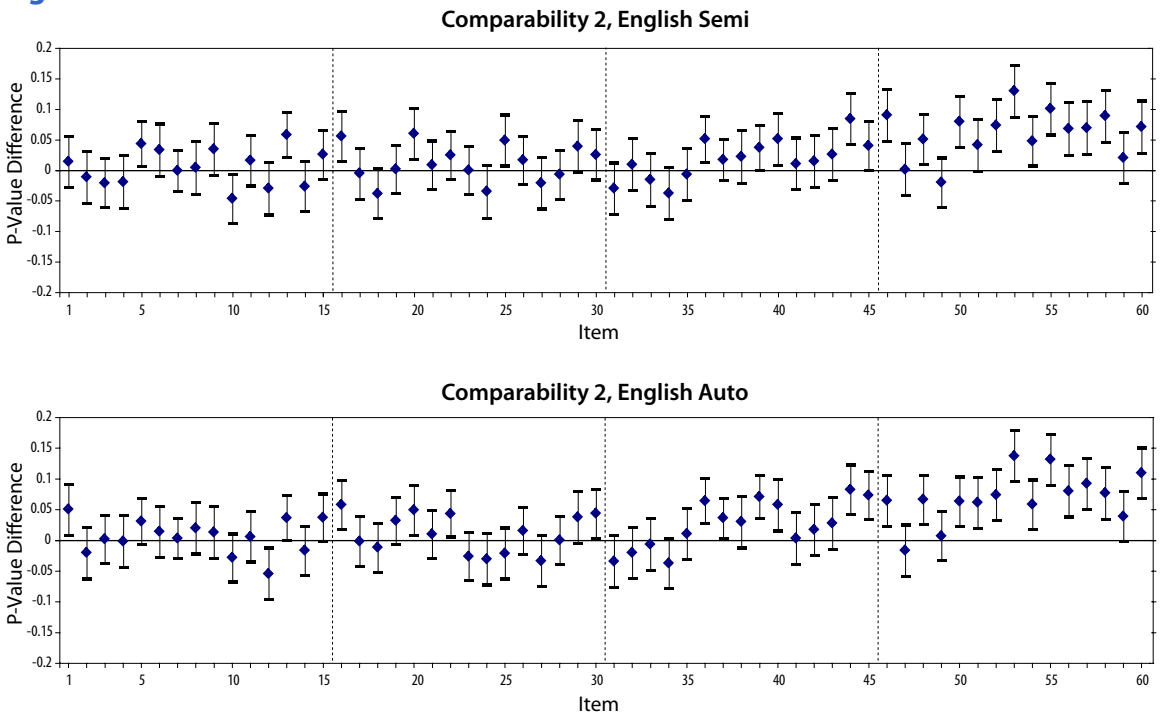

Figure 1: Computer–paper p-value differences ± 2 standard errors for Comparability 2 English test computer conditions (Semi and Auto).

Figure 2

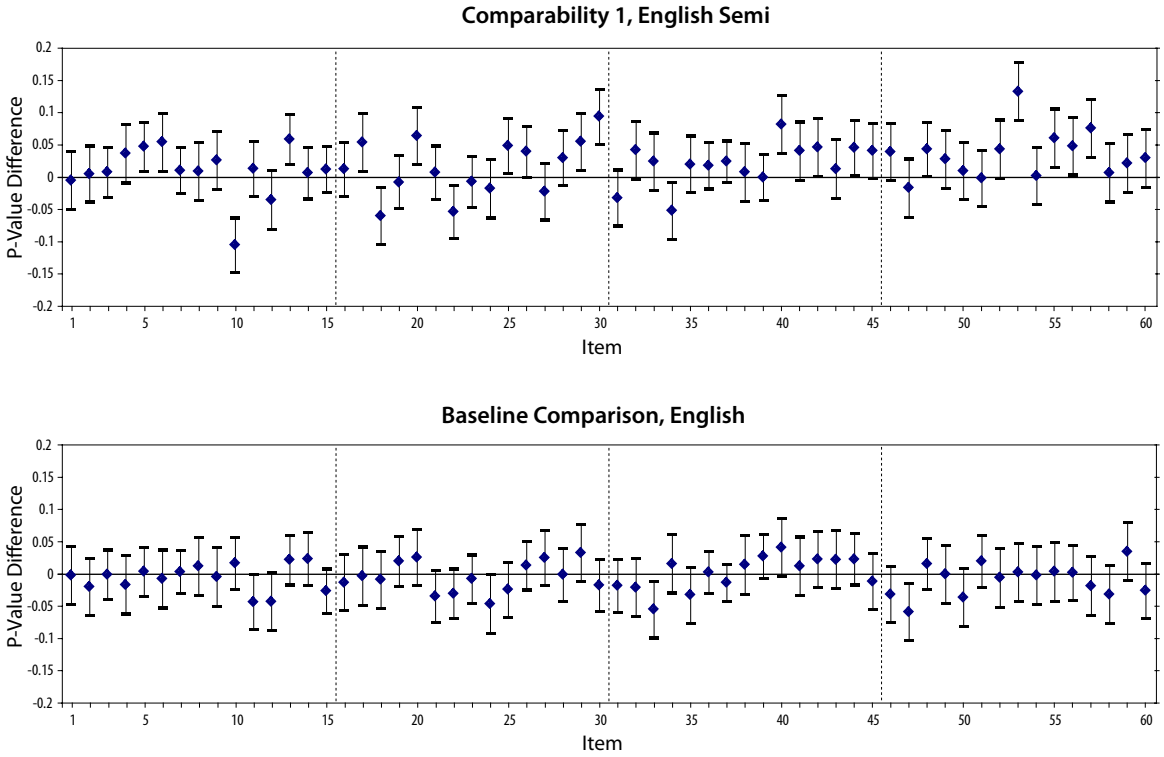

Figure 2: Computer–paper p-value differences ± 2 standard errors for Comparability 1 English test and the baseline comparison.

Figures 3–4 show similar results for Reading, while Figures 5–6 show similar results for Science Reasoning. In each of the plots in Figures 1–6, a positive difference indicates the item was easier on computer than on paper. We would expect that if there was no significant difference in performance across modes the error bands would surround zero (i.e., zero would not fall outside of the span of the p-value difference ± 2 standard errors).

(Figures 3, 4, 5 and 6 are shown on the following pages.)

Figure 3

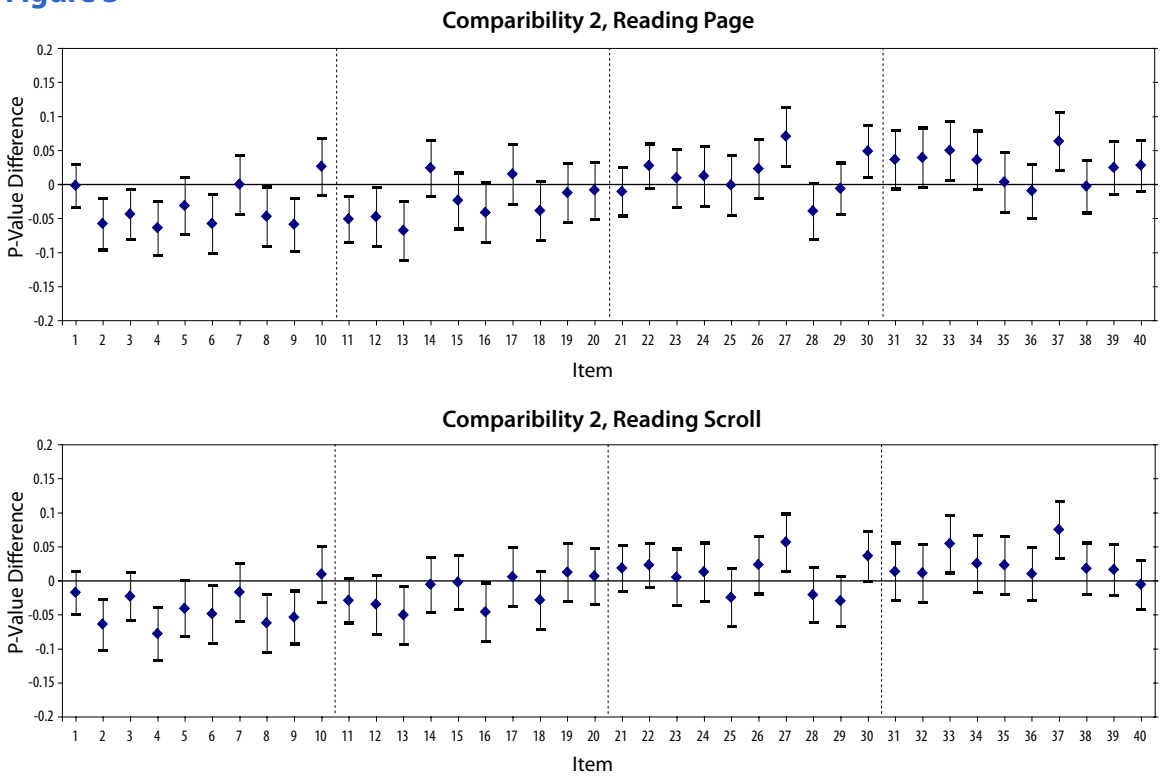

Figure 3: Computer–paper p-value differences ± 2 standard errors for Comparability 2 Reading test computer conditions (Page and Scroll).

Figure 4

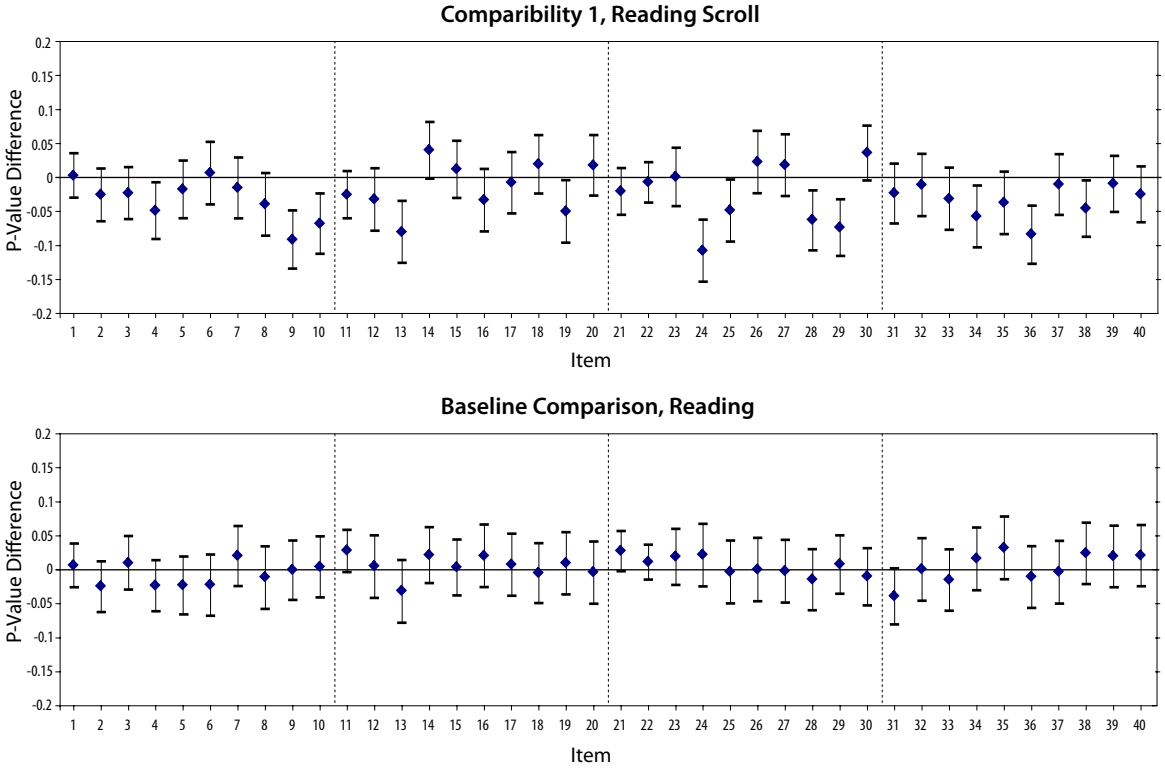

Figure 4: Computer–paper p-value differences ± 2 standard errors for Comparability 1 Reading test and the baseline comparison.

Figure 5

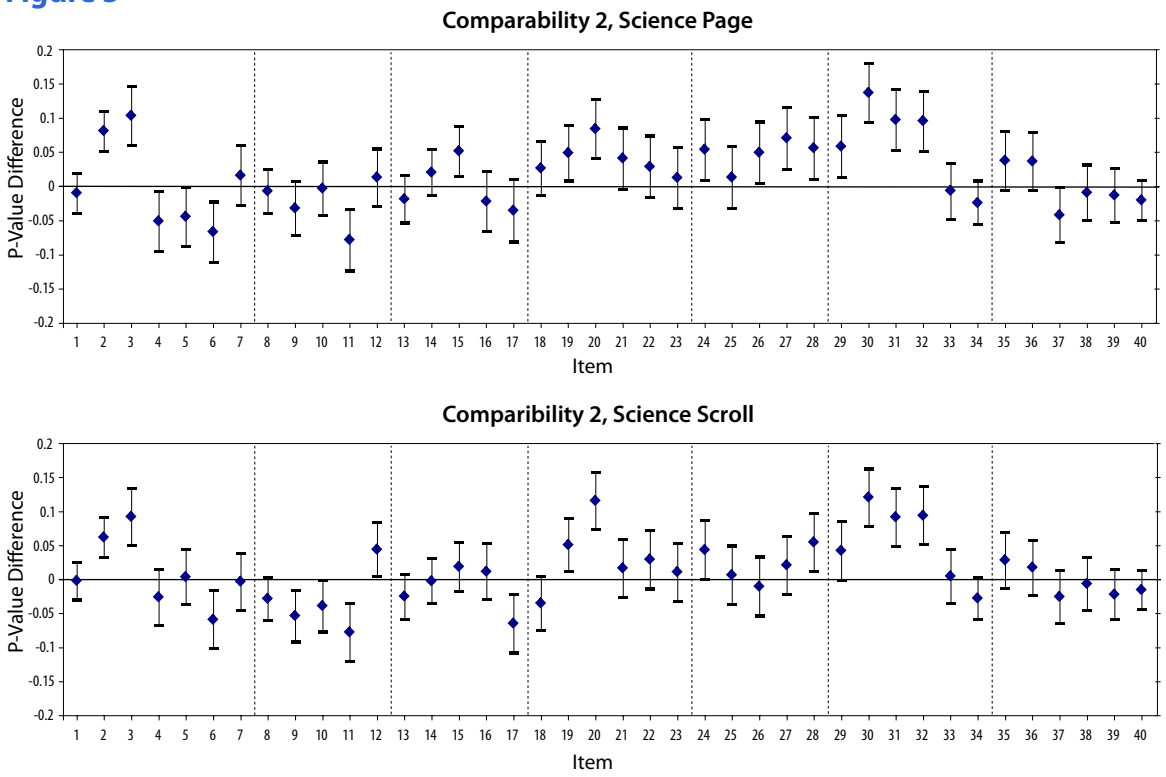

Figure 5: Computer–paper p-value differences ± 2 standard errors for Comparability 2 Science test computer conditions (Page and Scroll).

Figure 6

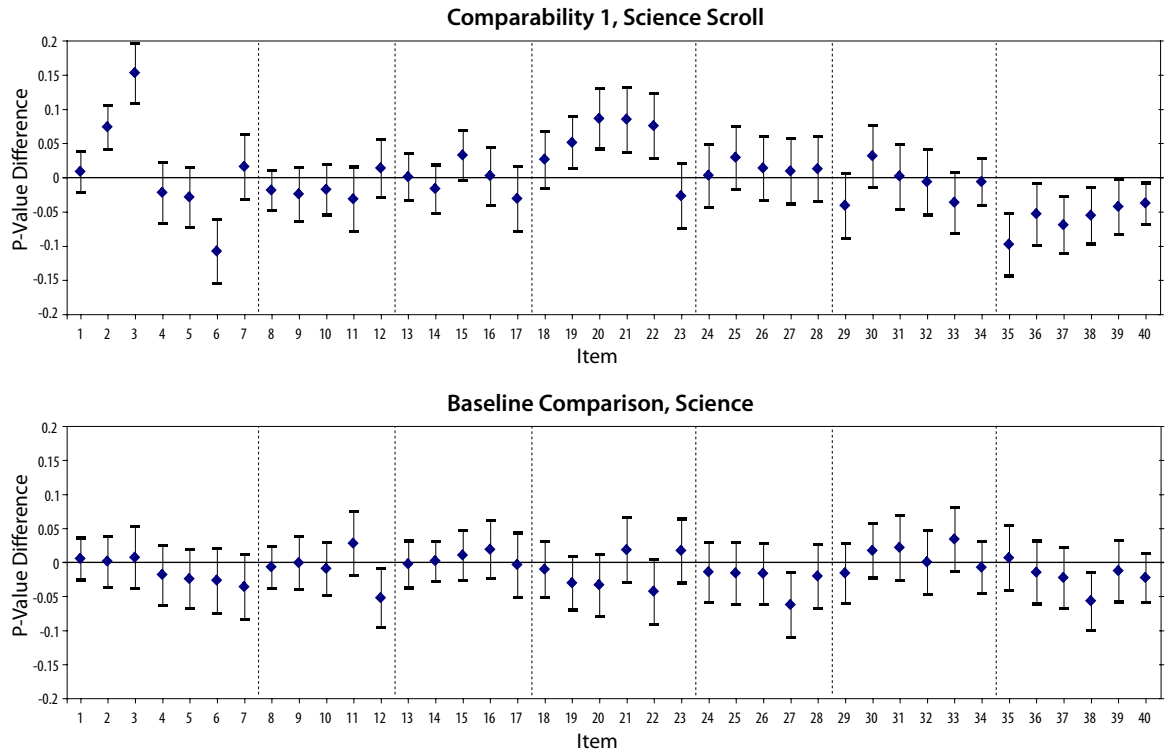

Figure 6: Computer–paper p-value differences ± 2 standard errors for Comparability 1 Science test and the baseline comparison.

Table 5 summarizes the number (and percent) of items showing significant p-value differences across modes in Comparability 1, Comparability 2, and for the baseline comparison, for all content areas. By chance alone, we would expect a small number of items from each test to display a significant performance difference. For example, at a 5% chance rate, we would expect three English, two Reading, and two Science Reasoning items to show significant differences. The baseline comparison provides some indication of how many significant differences we might realistically expect due to chance alone for each test, since the groups being compared took the same form via the same administration mode and were randomly equivalent. For each test, the total number of items showing significant differences was much larger for both Comparability 1 and Comparability 2 than for the baseline comparison. Thus, it appears that most of the significant differences observed across the two studies may be attributable to administration mode rather than chance. A few of the flagged differences may be spurious, but unfortunately, we cannot determine with any certainty which items show genuine differences, and which items show chance differences.

Table 5 Number (and Percent) of Items Showing Significant P-Value Differences Across Administration Modes

Test	Source	Condition	# Favoring Computer	# Favoring Paper
English	Comparability 1	Computer (Semi) vs. Paper	16 (27%)	4 (7%)
	Comparability 2	Semi vs. Paper	21 (35%)	1 (2%)
	Comparability 2	Auto vs. Paper	23 (38%)	1 (2%)
	Baseline	"Computer" vs. "Paper"	0 (0%)	4 (7%)
Reading	Comparability 1	Computer (Scroll) vs. Paper	0 (0%)	12 (30%)
	Comparability 2	Scroll vs. Paper	3 (8%)	8 (20%)
	Comparability 2	Page vs. Paper	4 (10%)	9 (23%)
	Baseline	"Computer" vs. "Paper"	0 (0%)	0 (0%)
Science Reasoning	Comparability 1	Computer (Scroll) vs. Paper	6 (15%)	7 (18%)
	Comparability 2	Scroll vs. Paper	10 (25%)	5 (13%)
	Comparability 2	Page vs. Paper	13 (33%)	5 (13%)
	Baseline	"Computer" vs. "Paper"	0 (0%)	3 (8%)

English Item Level Performance

Table 5 shows a large number of items significantly favoring computer examinees and only a handful of items significantly favoring paper examinees, for Comparability 1 and both the Auto and Semi conditions of Comparability 2. More items favored computer examinees in Comparability 2 than in Comparability 1. Fewer items favored paper examinees in Comparability 2 than in Comparability 1. Table 6 shows the number of items with significant differences for the first 35 and last 25 items of the test. Fewer items favored computer examinees in the beginning of the test for Comparability 2 than Comparability 1, and more items favored computer

examinees at the end of the test for Comparability 2 than for Comparability 1. The end-of-test trend favoring computer examinees in Comparability 2 is apparent in Figure 1.

Table 6 **Number (and Percent) of English Items Showing Significant P-Value Differences Across Administration Modes for the First 35 and Last 25 Items**

English Items	Source	Condition	# Favoring Computer	# Favoring Paper
First 35	Comparability 1	Computer (Semi) vs. Paper	8 (23%)	4 (11%)
	Comparability 2	Semi vs. Paper	5 (14%)	1 (3%)
	Comparability 2	Auto vs. Paper	5 (14%)	1 (3%)
Last 25	Comparability 1	Computer (Semi) vs. Paper	8 (32%)	0 (0%)
	Comparability 2	Semi vs. Paper	16 (64%)	0 (0%)
	Comparability 2	Auto vs. Paper	18 (72%)	0 (0%)

The interface changes (such as the removal of the full yellow highlighting) might have created more parity across paper and computer administrations early in the test, before speeded response behavior kicked in. English had the highest completion rates of all content areas in both Comparability 1 and Comparability 2, but the average amount of time spent on each item for computer examinees suggests that there was some speeded response behavior later in the test. On average, Semi examinees spent 84.4 seconds per item on the first 35 items (83.2 seconds for Auto). For the last 25 items, however, Semi examinees spent, on average, only 25.8 seconds per item (25.6 seconds for Auto).

Once examinees start to rush to complete the exam, it might be advantageous to take the test on computer rather than on paper, because of the ease in responding and moving quickly through items, and a greater ability to focus on the item at hand without being distracted by extraneous information. In discussions with examinees that expressed a preference testing on computer, many mentioned that they preferred testing on computer because it was easier not having to bubble in the answers. The hypothesized ease of engaging in speeded response behavior on the computer relative to paper will subsequently be referred to as the "no-bubble effect." The type of speeded response behavior hypothesized here should not be confused with purely random responding that might also occur at the very end of a test, whereby examinees fill in random responses without reading the question at all.

Reading Item Level Performance

Table 5 shows that roughly a fourth of the test items significantly favored paper examinees in Comparability 1 and in both conditions of Comparability 2. Fewer items favored paper examinees in both conditions of Comparability 2 than in Comparability 1. A handful of items significantly favored computer examinees for both

the Scroll and Page conditions in Comparability 2. No items favored computer examinees in Comparability 1. Table 7 shows the number of items showing significant differences for the first 13 and last 14 items of the test. More items favored paper examinees at the beginning of the test for both conditions of Comparability 2, than Comparability 1. Fewer items favored paper examinees and more items favored computer examinees at the end of the test, for Comparability 2 than Comparability 1. The beginning-of-test trend favoring paper examinees in Comparability 2 is apparent in Figure 3.

Table 7 **Number (and Percent) of Reading Items Showing Significant P-Value Differences Across Administration Modes for the First 13 and Last 14 Items**

Reading Items	Source	Condition	# Favoring Computer	# Favoring Paper
First 13	Comparability 1	Computer (Scroll) vs. Paper	0 (0%)	4 (31%)
	Comparability 2	Scroll vs. Paper	0 (0%)	7 (54%)
	Comparability 2	Page vs. Paper	0 (0%)	9 (69%)
Last 14	Comparability 1	Computer (Scroll) vs. Paper	0 (0%)	5 (36%)
	Comparability 2	Scroll vs. Paper	3 (21%)	0 (0%)
	Comparability 2	Page vs. Paper	4 (29%)	0 (0%)

The shift observed in Comparability 2 from favoring paper examinees at the beginning of the test to favoring computer examinees toward the end of the test might be attributable to the no-bubble effect. Results for Comparability 2 suggest there was some speeded response behavior. On average, Page examinees spent 51.3 seconds per item on the first 15 items, and 31.6 seconds per average on the last 15 items (51.6 and 31.6 seconds for Scroll examinees on the first 15 items and last 15 items, respectively). The favoring of paper examinees at the end of the test in Comparability 1 could have occurred because slower navigational speed interfered with the ability of computer examinees to move quickly through passages and items once speeded response behavior began.

At face value, it is unclear why more items favored paper examinees at the beginning of the test under Interface 2 than under Interface 1. The changes made to the interface were designed to improve the speed with which examinees could navigate throughout the passage (in the Scroll condition), to facilitate the occurrence of positional memory for specific content in the passage (in the Page condition), and to improve pre-test training on how to navigate. More items favoring paper examinees in the beginning was an unexpected outcome of the interface changes.

One possible explanation for this finding is that a slightly longer line length was used in Interface 1 than in Interface 2. (All other characteristics such as font size, spacing between lines, and the size of the passage window were the same across the two interfaces.) There was typically one more word per line in Interface 1

than Interface 2. This resulted in more lines per passage in Interface 2 than Interface 1. In the first passage of the test, the different line lengths resulted in 81 total lines of text in Interface 1 versus 89 total lines of text in Interface 2. Research has suggested that reading speed for scrolled text increases with increases in line length (Duchnicky & Kolers, 1983), and that reading speed decreases as the number of words per page decreases (Muter, 1996), so it is possible that the different line lengths differentially affected the reading speed of computer examinees across the two studies. The additional number of lines in Interface 2 also meant that more navigation was required than in Interface 1. If examinees were still learning how to navigate early in the test, the extra navigation required in Interface 2 could also have been a factor that caused more items early in the test to favor paper examinees under Interface 2. This effect could have been counteracted later in the test by the onset of speeded response behavior.

Science Reasoning Item Level Performance

Table 5 shows no clear direction of favoritism for Comparability 1. A similarly moderate number of items showed significant differences for both computer and paper examinees. For both the Page and Scroll conditions of Comparability 2, more items favored computer examinees than paper examinees. More items favored computer examinees in both conditions of Comparability 2 than in Comparability 1. Slightly fewer items favored paper examinees in both conditions of Comparability 2 than Comparability 1.

Table 8 shows the number of items showing significant differences for selected passages. In Comparability 1, some trends in performance differences occurred within certain passages of the test. Items in Passage 4 strongly favored computer examinees, whereas all items in the last passage (Passage 7) significantly favored paper examinees. These trends are apparent in Figure 6. Within the remaining passages, there was no apparent trend favoring either computer or paper examinees. The effect for the last passage might be attributable to speeded response behavior and slow navigational capabilities, which could have disadvantaged computer examinees. Results for Comparability 2 suggest there was still some speeded response behavior under Interface 2. On average, Page examinees spent 54.8 seconds on the first 15 items, and 30.3 seconds on the last 15 items (56.3 and 30.9 seconds for Scroll examinees on the first 15 items and last 15 items, respectively).

For Comparability 2, findings were a bit different within passages. The last passage (Passage 7) was neutral for both the Page and Scroll conditions (with the exception of Item 37 significantly favoring paper examinees in the Page condition). This suggests that the interface and tutorial changes might have removed some of the factors that caused computer examinees to be disadvantaged in Comparability 1 once speeded response behavior began. Passage 4 showed slightly fewer items favoring computer examinees than in Comparability 1, for both the Page and Scroll conditions. Within Passages 5–6, there was a trend for items to favor computer examinees that did not occur in Comparability 1. The trend occurred more so in the Page condition than the Scroll condition (see Figure 5).

Table 8 Number (and Percent) of Science Reasoning Items Showing Significant P-Value Differences Across Administration Modes for Selected Passages

Science Reasoning Items	Source	Condition	# Favoring Computer	# Favoring Paper
Passages 1–2 (12 items)	Comparability 1	Computer (Scroll) vs. Paper	2 (17%)	1 (8%)
	Comparability 2	Scroll vs. Paper	3 (25%)	4 (33%)
	Comparability 2	Page vs. Paper	2 (17%)	4 (33%)
Passage 4 (6 items)	Comparability 1	Computer (Scroll) vs. Paper	4 (67%)	0 (0%)
	Comparability 2	Scroll vs. Paper	2 (33%)	0 (0%)
	Comparability 2	Page vs. Paper	2 (33%)	0 (0%)
Passages 5–6 (11 items)	Comparability 1	Computer (Scroll) vs. Paper	0 (0%)	0 (0%)
	Comparability 2	Scroll vs. Paper	5 (45%)	0 (0%)
	Comparability 2	Page vs. Paper	8 (73%)	0 (0%)
Passage 7 (6 items)	Comparability 1	Computer (Scroll) vs. Paper	0 (0%)	6 (100%)
	Comparability 2	Scroll vs. Paper	0 (0%)	0 (0%)
	Comparability 2	Page vs. Paper	0 (0%)	1 (17%)

The trend in the middle-to-end of the test favoring computer examinees in Comparability 2 might be attributable to the "focus effect." It might be beneficial for computer examinees to be able to view only the relevant graphic on screen, with extraneous information contained in the rest of the test not visible on screen. It is unclear, however, why the last passage did not favor computer examinees in Comparability 2, as was observed for English and Reading. Since the completion rates were similar for computer examinees across the Reading and Science Reasoning tests, which contained the same number of items, we would expect a similar end-of-test effect across the two tests. Having to compare information across tables and figures in Science Reasoning may offset the no-bubble effect to some degree.

Across the first two passages, there appeared more of a trend for items to favor paper examinees in Comparability 2 than in Comparability 1. A similar trend was noted for the Reading test. That similar trends favoring paper examinees were observed in the beginning of the test for both Reading and Science Reasoning under Interface 2 suggests the same factors could be contributing to those results across the two tests. Although some items early on favored paper examinees, Items 2 and 3 significantly favored computer examinees in Comparability 1 and Comparability 2. What caused this performance difference in Items 2 and 3 across both studies is unclear. There were possibly some content factors that caused them to strongly favor computer examinees.

Item Analysis

Detailed evaluations of item and interface design features were conducted for selected English and Reading items in an attempt to identify some of the different factors that could have contributed to item-level performance differences across administration modes. Hypotheses to account for performance differences observed in Comparability 1 were originally developed for these items by test specialists after a review of the test content, test booklet, interface features, and interviews with examinees (see also Pommerich & Burden, 2000). Interface changes were then made between Comparability 1 and Comparability 2 in response to these hypotheses. The hypotheses are re-evaluated here, using the findings across the two comparability studies for the selected items. Some possible explanations for the findings across studies are offered, given the item characteristics and interface features. These explanations are only speculations, because it cannot be known with any certainty from these studies exactly what caused the performance differences. Also, be reminded that differences for some individual items may be due to chance alone.

Figure 7 shows p-value difference plots for English Items 6, 10, 13, 17, 18, 22, and 30, while Figure 8 shows p-value difference plots for Reading Items 4, 6, 9, 24, 25, 29, and 30. Each item plot contains the computer – paper p-value differences ± 2 standard errors for Comparability 1, the two conditions in Comparability 2, and the baseline comparison. The hypotheses that were posited to explain Comparability 1 results for the seven selected English items focused on the use of highlighting in the computer presentation, different alignments of the underlined portions with the corresponding item across modes, and different passage layouts across modes. The hypotheses that were posited to explain Comparability 1 results for the seven selected Reading items focused on different line breaks and passage layouts across modes, and the navigation required to answer the item. Although some general factors were identified across the selected items in response to Comparability 1 findings, further analysis of Comparability 2 results suggests that a combination of factors may have contributed to differential performance across modes.

Figure 7

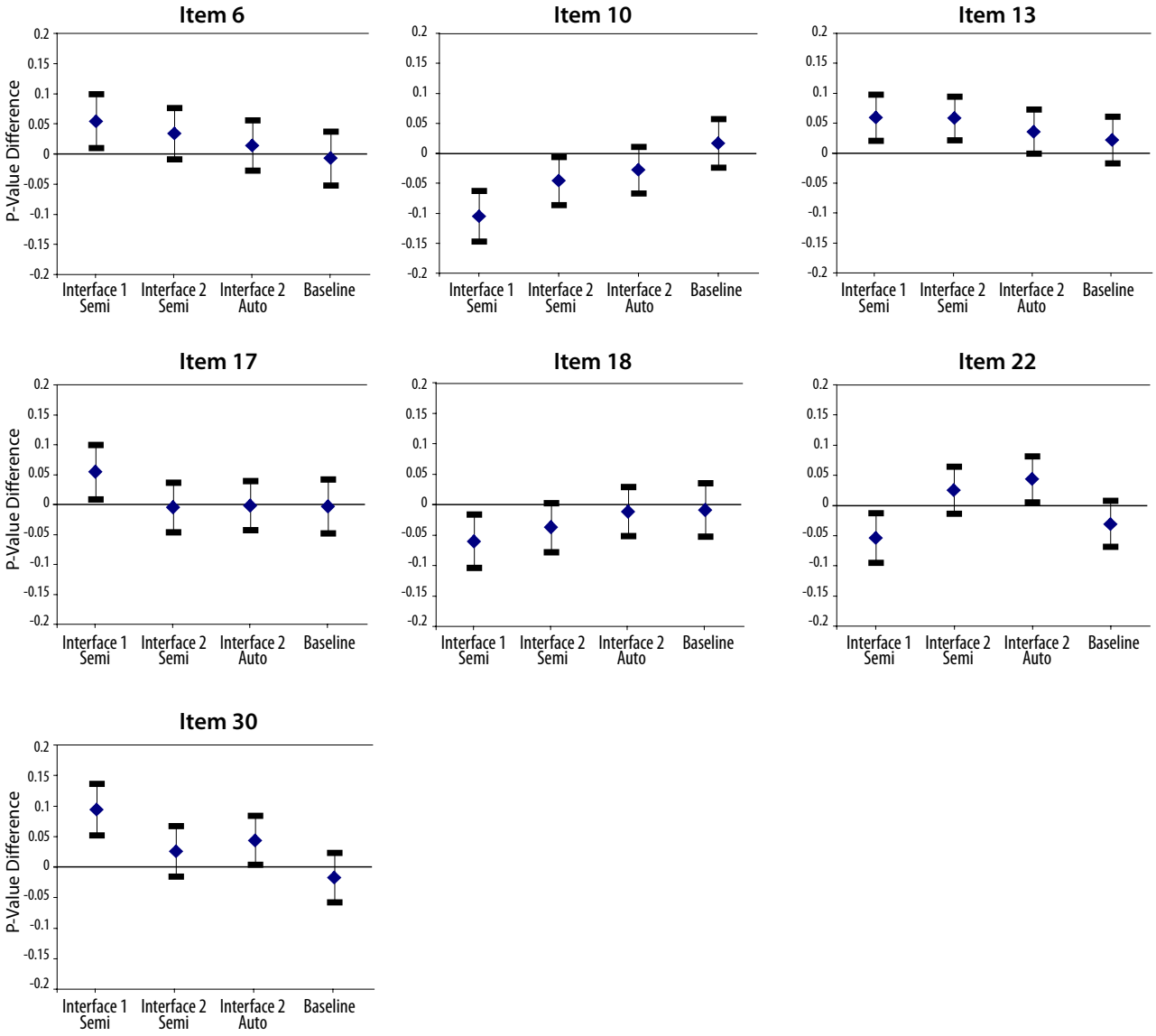

Figure 7: Computer–paper p-value differences ± 2 standard errors for select English items for Comparability 1 (Interface 1-Semi), the two computer conditions in Comparability 2 (Interface 2-Semi and Interface 2-Auto), and the baseline comparison.

Figure 8

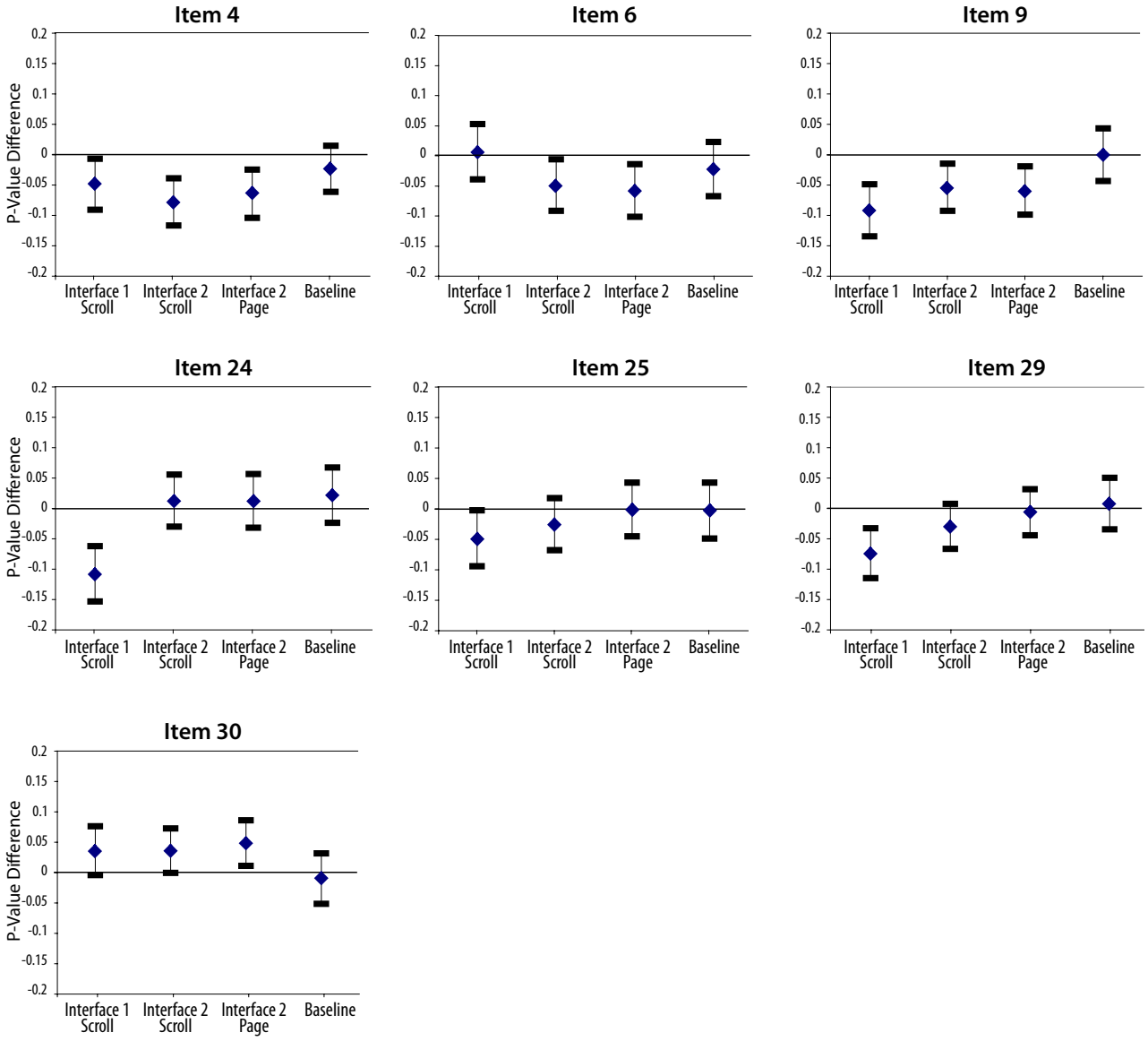

Figure 8: Computer–paper p-value differences ± 2 standard errors for select Reading items for Comparability 1 (Interface 1-Scroll), the two computer conditions in Comparability 2 (Interface 2-Scroll and Interface 2-Page), and the baseline comparison.

Factors hypothesized as contributing to performance differences are discussed as follows:

- highlighting of text;
- alignment of item with text;
- layout of passage;
- location of line breaks;
- ease of navigation.

Highlighting

For some items under Interface 1, it was hypothesized that the color highlighting of the full underlined portion helped computer examinees focus on the full underlined portion and read it in its entirety, which enabled them to more easily identify correct or incorrect responses. Underlining and the use of color are each techniques that can be used alone to highlight and draw attention (Tullis, 1997). Fisher and Tan (1989) recommend the use of color to highlight whenever possible. The use of two cues (underlining and color) in the computer presentation versus one cue (underlining) in the booklet presentation may have had different effects across modes in drawing attention to the relevant portion. Findings for English Items 6 and 17 will be discussed with respect to the highlighting hypothesis. Both items significantly favored computer examinees in Comparability 1.

English Item 17. The underlined portion for this item contained a word that might have been unfamiliar to many examinees. The underlined portion was grammatically correct as written, and thus "No Change" was the correct response. It was hypothesized that the color highlighting helped focus attention on the underlined portion, so that Comparability 1 computer examinees unfamiliar with the word may have been more inclined to consider the underlined portion as a viable response option and thereby choose "No Change." In Interface 2, the full color highlighting was removed and only the item number underneath the underlined portion was highlighted with color. The results for both the Auto and Semi condition showed no favoritism for computer examinees under the new interface.

English Item 6. In the test booklet for the paper administration, a page break occurred in the middle of the sentence that contained the underlined portion for this item. In Comparability 1, a greater percentage of paper examinees (19.4%) selected the incorrect option A, "No Change," than did computer examinees (12.8%). It was hypothesized that the color highlighting of the full underlined portion in Interface 1 might have helped some computer examinees better focus on the underlined portion in the context of the complete sentence and recognize that "No Change" was not a valid option, whereas paper examinees might have been inclined to incorrectly respond "No Change" if they did not consider the complete sentence containing the underlined portion in responding.

In Interface 2, results for the Semi condition favored computer examinees, although the difference was not significant, while results for the Auto condition showed an even smaller difference. Approximately 20.6% of Comparability 2 paper examinees selected Option A, versus 18.5% of Comparability 2 Semi examinees and 19.2% of Comparability 2 Auto examinees. A greater percentage of computer examinees selected Option A in the Auto and Semi conditions than in Interface 1 (12.8%). The removal of the full color highlighting, along with the presentation of an item number adjacent to Option A may have made Semi examinees more inclined to select Option A in Interface 2 than in Interface 1. Auto examinees performed similarly to paper examinees in Interface 2. They may have been further inclined to select Option A than Semi examinees because the underlined portion

was aligned with Option A. The alignment of the underlined portion with the item in the Interface 2 Auto condition matched that of the paper presentation, while the alignment in the Interface 2 Semi condition matched that of the Interface 1 presentation (which differed from the paper presentation).

Alignment

As results for English Item 6 suggest, differing alignments of the underlined portion with the item may differentially influence examinee behavior. For some items under Interface 1, it was hypothesized that where the underlined portion was aligned with the item had an effect on examinees' behavior. Findings for English Items 10, 13, 18, and 22 will be discussed with respect to the alignment hypothesis. English Item 13 significantly favored computer examinees in Comparability 1, while English Items 10, 18, and 22 all significantly favored paper examinees in Comparability 1. Items 10, 18, and 22 all contained a stimulus that asked a question about the underlined portion in the passage, followed by the response options. Item 13 contained no stimulus (i.e., only response options were listed for the underlined portion).

No Stimulus

English Item 13. All response options for this item looked acceptable. It was hypothesized that the alignment of response options with the underlined portion in Interface 1 may have influenced the Comparability 1 computer examinees' responses. The correct answer was D, and the underlined portion was aligned below Option D. In interviews with examinees (described in Pommerich & Burden, 2000), some computer examinees indicated reading the response options from bottom to top for this item. Thus, Comparability 1 computer examinees may have been more inclined to read the response options from bottom to top and select the first acceptable-looking response (i.e., D). In Interface 2, the underlined portion was also aligned below Option D under the Semi condition, whereas for the Auto condition, the underlined portion was aligned with the top of the response options (i.e., with Option A). The results for the Semi condition in Comparability 2 were similar to results for Comparability 1, and significantly favored the computer examinees. The results for the Auto condition also favored the computer examinees, but the difference was not significant. It is possible that when the underlined portion falls below the response options, examinees might be more inclined to read the options from bottom to top. When the underlined portion is aligned with the top of the item, examinees might be more inclined to read the options from top to bottom.

Stimulus

English Item 10. This was the first item on the test that contained a stimulus that asked a question about the underlined portion in the passage. Examinees needed to read the stimulus to understand how to respond to the item. It was hypothesized that Comparability 1 computer examinees were less likely to read the stimulus than paper examinees because the underlined portion was not aligned with the top of

the item (i.e., with the stimulus), and there was no numbering of the item to draw attention to the stimulus. In Interface 2, item numbering was added adjacent to the top of the item for both the Auto and Semi conditions. It was hoped that the item number would help draw examinees' attention to the stimulus. For the Semi condition, the underlined portion was not aligned with the top of the stimulus, whereas for the Auto condition, the underlined portion was aligned with the top of the stimulus. Results for Comparability 2 still significantly favored paper examinees for the Semi condition, but to a lesser degree than in Comparability 1. Thus, it appears that the inclusion of the item number may have helped draw some Semi examinees' attention to the stimulus, but that the lack of alignment of the underlined portion with the stimulus could still have caused some computer examinees to ignore the stimulus while responding. Results for the Auto condition, which did align the underlined portion with the stimulus, also favored paper examinees, but the difference was not significant.

English Item 18. This item also contained a stimulus that examinees had to read to answer. It was difficult to guess correctly if the stimulus was not read. As with Item 10, it was hypothesized that Comparability 1 computer examinees were less likely to read the stimulus than paper examinees because the underlined portion fell below the stimulus and there was a lack of focus on the stimulus. Results for this item were fairly similar to results for Item 10. The results for the Interface 2 Semi condition favored paper examinees, but not significantly. The results for the Auto condition did not favor either group. As in Item 10, whether the underlined portion was aligned with the stimulus appeared to have a bigger influence on computer examinees' tendency to read the stimulus than the inclusion of the item number, although the item numbering may have helped somewhat.

English Item 22. This item contained a stimulus that referred to the previous two sentences of the passage and asked examinees to select the response option that paralleled the style used in those sentences. The style to follow was not apparent without reading the previous two sentences. Because the item contained a stimulus, we expected similar results as observed for Items 10 and 18. The Comparability 2 results, however, were different for this item than for Items 10 and 18. The results for the Semi condition favored computer examinees, but not significantly. The results for the Auto condition significantly favored computer examinees.

One plausible explanation for the Comparability 2 findings has to do with the different layout of the relevant sentences across the different interfaces. Auto examinees needed to scroll up to see both of the referenced sentences, and so may have been more likely to actually read the referenced sentences than paper examinees, if they read the stimulus. The results suggest that the alignment of the underlined portion with the stimulus in the Auto condition may have influenced computer examinees to read the stimulus, and thereby to scroll and read the referenced sentences. Among paper and Auto examinees that read the stimulus, Auto examinees may have read the relevant sentences more carefully than the paper examinees because they had to scroll to see them.

In the Semi condition, although the underlined portion occurred on the same line in the passage as in Interface 1, the width of the passage window was slightly smaller in Interface 2, so that each line contained fewer words and the line breaks differed across the two interfaces. Thus, the underlined portion occurred at the beginning of the line in the Semi condition versus in the middle of the line in Interface 1, and the layout of the two preceding sentences differed. Given the different layouts and/or the fewer words per line in the Semi condition, it is possible that the style used in the two referenced sentences was more obvious in the Interface 2 Semi condition than in Interface 1. Because of the similar lack of alignment of the underlined portion with the stimulus, Interface 2 Semi examinees might not have been any more likely to read the stimulus than Interface 1 examinees, but it could have been easier for them to infer the correct answer without reading the stimulus because of the layout of the relevant sentences.

Passage Layout

As results for English Item 22 suggest, differing passage layouts for an item may differentially affect examinee behavior. For some items under Interface 1, it was hypothesized that computer examinees were advantaged by the passage layout and the information that was visible on screen when the item was selected. Findings for English Item 30 and Reading Item 30 will be discussed with respect to the passage layout hypothesis. English Item 30 significantly favored computer examinees in Comparability 1. Reading Item 30 also favored computer examinees in Comparability 1, although not significantly.

English Item 30. This item required the examinee to choose the correct tense for an underlined word. The underlined portion for this item was contained in the last paragraph in the passage. It was hypothesized that the passage layout in the booklet versus on screen influenced performance in Comparability 1. In the test booklet, the last paragraph appeared alone on a page, whereas in Interface 1, the last two paragraphs of the passage were visible on screen when this item was selected. It is likely that with more of the passage visible, it was easier for computer examinees to correctly infer the tense of the passage.

In Comparability 2, results for the Semi condition did not significantly favor either group. In the Semi condition, only the last paragraph was visible on screen when this item was selected, which matched the page layout for this paragraph in the test booklet. The last passage contained the underlined portions for Items 28–30. As discussed earlier, although the Semi condition scrolled at the same time as Interface 1 in Comparability 1, the width of the passage window and number of words per line differed across Interface 1 and 2, leading to different parts of the passage being visible for the same item. Results for Semi were similar to the paper results, likely due to the similar layout on screen.

Results for the Auto condition significantly favored computer examinees, although to a lesser degree than computer examinees were favored in Comparability 1. In the Auto condition, only a portion of the last paragraph was visible on screen when this item was selected (six out of eight total lines). These six lines

contained the underlined portions for Items 29 and 30. Item 29 also required the examinee to choose the correct tense for an underlined word, and the correct response was "No Change," so the underlined portion for Item 29 reads correctly as is in the passage. It is possible that Auto computer examinees were more likely to respond correctly to Item 30 than paper examinees because it was easier to infer the correct tense from the information visible on screen. The slight difference in the passage layout across the Auto and booklet presentations appears to have had an effect on responses.

Reading Item 30. This item was a very difficult item that required a global understanding of the passage. No explicit answer was stated in the passage. If performance on this item was consistent with performance on other items with similar characteristics, we would expect that it would have favored paper examinees because the answer was not explicitly stated and the examinee had to navigate to find the relevant information to answer the question. (The issue of navigation will be discussed more fully shortly.) However, it was hypothesized that computer examinees were advantaged by the passage layout on screen and their response to the previous question. The correct response for this item referred to the "blues" and was the only response option to contain the word blues. The paragraph that contained the answer to Item 29 also referred to the blues. Computer examinees that had the paragraph referring to the blues visible on screen from answering Item 29 might have been more inclined to select the response option for Item 30 that also referred to the blues, because the word blues was visible in the passage window. Results for Comparability 2 showed the same trend as Comparability 1, for both the Scroll and Page condition. Computer examinees were favored in both conditions, and the difference was significant for the Page condition.

Line Breaks

The use of different line breaks in passages across paper and computer modes may also contribute to mode effects, particularly for items that contain references to specific lines in the passage. For Reading Item 24 under Interface 1, it was hypothesized that the slightly different content of the referenced line (caused by different line breaks across paper and computer modes) caused a performance difference. Results for this item significantly favored paper examinees.

Reading Item 24. This item referred examinees to a specific line in the passage and asked the meaning of the term "blue" in the referenced line. The referenced line in the booklet presentation contained only the word "blue," whereas the referenced line in the Interface 1 computer presentation contained both the word "blue" and "blues," which could have been confusing to computer examinees. In Interface 2, the line breaks were identical across computer and paper modes, so the content of referenced lines was identical. We expected that there would be no performance difference for this item in Comparability 2, and there was not for either the Page or Scroll condition.

Navigation

The reading items typically required some amount of navigation throughout the passage in order to find relevant sections of the passage. For some items under Interface 1, it was hypothesized that slow navigation speed and lack of knowledge of navigation capabilities negatively affected the computer examinees' performance. Findings for Reading Items 4, 6, 9, 25, and 29 will all be discussed with respect to the navigation hypothesis. Reading Items 4, 9, 25, and 29 all significantly favored paper examinees in Comparability 1. Reading Item 6 did not favor either paper or computer examinees in Comparability 1, and was expected to be neutral in Comparability 2 also. Item 25 contained a line reference, while Items 4, 6, 9, and 29 contained no line reference. The answer was explicitly stated in the passage for Items 4 and 29, while the answer was not explicitly stated in the passage for Items 6 or 9.

Line Reference/Answer Explicitly Stated

Reading Item 25. This item referred examinees to a specific line in the passage. Item 24 also referred examinees to a specific line elsewhere in the passage, so that all examinees answering Items 24 and 25 in order had to move from the line referenced in Item 24 to the line referenced in Item 25. It was hypothesized that navigation difficulties and slow scrolling speed interfered with computer examinees' performance on this item. In the Scroll condition for Interface 2, scrolling speed was increased, and the pre-test training on scrolling was improved. A separate Page condition was added for comparison purposes. Results for Comparability 2 showed no significant difference in performance across paper and computer modes, for either the Scroll or Page condition, although the Scroll condition did show an advantage for paper examinees. Navigation may have been somewhat easier for Page examinees than for Scroll examinees.

No Line Reference/Answer Explicitly Stated

Reading Item 29. This item was a very difficult item. The answer was contained in the second to last paragraph of the passage, and no line reference was provided for this item. It was hypothesized that this item required a lot of navigation to find the answer in the passage, and that Comparability 1 computer examinees had difficulty navigating through the passage. It was expected that mode differences would be diminished under the improved navigation in both the Scroll and Page conditions in Comparability 2. Results for both the Scroll and Page conditions showed no significant difference, although the Scroll condition did show an advantage for paper examinees. Again, navigation may have been somewhat easier for Page examinees than Scroll examinees. Positional memory may also have been more likely to occur for Page examinees.

Reading Item 4. This item referred to a specific part of the passage, but no line reference was given, so there was undirected scrolling or paging to find the information in the passage. It was hypothesized that the item required a lot of scrolling to find the appropriate section of the passage and that Comparability 1 examinees

were hampered by slow navigation speed. It was expected that similar to Item 29, mode differences would be diminished in Comparability 2. However, the results for both the Scroll and Page conditions significantly favored paper examinees.

The percentages responding to each option for Item 4 show that computer examinees in Comparability 1 and both conditions of Comparability 2 were distracted by another response option that contained information that was given in the passage, but that was not the correct response to the question. It is possible that computer examinees were more likely to stop reviewing the passage upon identifying a correct-looking option than were paper examinees. The extra navigation required to evaluate all response options may have prohibited computer examinees from continued checking after selecting an initial correct-looking response. This tendency may also have been more likely early in the test if examinees were still familiarizing themselves with the navigational capabilities. It appears that improving the navigational capabilities in Interface 2 did not have the intended effect for this item.

No Line Reference/Answer Not Explicitly Stated

Reading Item 9. This item was a difficult item. The answer was not explicitly stated in the passage and required undirected scrolling to find relevant information in the passage. Again, it was expected that improved navigation in Interface 2 would decrease the mode effect that was observed in Comparability 1. However, the item still significantly favored paper examinees in Comparability 2, for both the Scroll and Page conditions. As with Reading Item 4, the computer examinees in both studies were distracted by one reasonable looking response option.

Reading Item 6. This item assumed a global understanding of the passage. The answer was not stated directly in the passage, but rather, the reader had to infer from the passage the correct response. Results for Comparability 1 did not favor either paper or computer examinees, and the item was expected to perform similarly in Comparability 2. However, results for both the Scroll and Page condition in Comparability 2 significantly favored paper examinees. On average, the Comparability 2 computer examinees did not spend any more time on this item than other items, so it does not appear that they were looking for an answer that they could not find. They did, however, choose an incorrect answer more frequently than paper examinees did. As with Reading Items 4 and 9, the computer examinees in Comparability 2 appeared distracted by a reasonable looking option that could be inferred as correct by someone who did not read the passage carefully. It is not clear why Comparability 2 computer examinees were negatively affected while Comparability 1 examinees were not. The faster navigation capabilities of Interface 2 plus the additional number of lines to be navigated (due to shorter line lengths) may have resulted in examinees reviewing the passage less carefully for this item under Interface 2.

Discussion

The findings from the two comparability studies, in conjunction with experience garnered from reviews of test content, test booklets, computer interfaces, and interviews with examinees, suggest some answers to the two questions broached in this paper. While some items showed no significant performance differences across administration modes, there were other items for which examinees clearly did not respond in the same way across modes or interface variations. The evaluation of the individual English and Reading items suggests that there are a variety of factors that could contribute to mode effects, and that each item presents a potentially unique set of circumstances that could cause different (and unpredictable) behaviors across modes. The results appeared to be affected by the different characteristics of each test and the position of the items in the test. The findings suggest that we should not expect the same relative performance across modes at the beginning and end of a given test, and that we should not expect the same relative performance across modes at the same point in the test (i.e., the beginning or end) for different tests.

Intuition suggests that the more complex the test is, and the greater the differences in how passages and items are presented across modes, the greater the potential for performance differences across modes. For complex tests where the information for an item cannot be displayed on screen all at once, it is probably not possible to develop a computer interface that would eliminate mode effects completely. However, by paying careful attention to examinees' test-taking practices, it may be possible to design test booklets and computer interfaces such as to minimize mode effects.

There may have been some sampling differences and some chance differences that affected the results of these studies, but in general, it appears that the changes made to the interface between the comparability studies had some effect on computer examinees' performance on some items. For some items, the effect was the intended effect, but for other items, the effect was not the intended effect. These findings suggest that examinees are sensitive and respond to how information is presented on computer, but not always in ways that are readily predictable. In some cases, the results appeared influenced by better pre-test training on how to use the functions necessary to take the test on computer, improved navigation, and more readily available information about the test session. While perhaps not all technically part of the computer "interface," each of these components contributes to the examinees' interaction with the interface, and should be considered in designing an interface and conducting computerized testing.

Different results across interface variations in Comparability 2 also suggest that even within the same mode of administration, differences in how the test is presented could influence examinee behavior while testing. Although there were no overly compelling differences in performance across the interface variations studied (scrolling versus paging in Reading and Science Reasoning, automatic scrolling versus semi-automatic scrolling for English), examinees responded

differently to some items under the different variations. For both Reading and Science Reasoning, average scores and completion rates were slightly higher for the Page condition than for the Scroll condition, and more items favored computer examinees in the Page condition. Strong preferences have been stated for paging over scrolling by some researchers (Schwarz, Beldie, & Pastoor, 1983; Dillon, 1992; Muter, 1996). For English, average scores and completion rates were slightly higher for the Auto condition than the Semi condition, and more items favored computer examinees in the Auto condition. Results for individual English items with a stimulus remind us that a seemingly subtle change such as aligning or not aligning the underlined portion in the English test with the top of the item can have a not-so-subtle effect on examinee behavior on some items. Thus, care also needs to be taken when implementing interface changes in an operational computerized testing program.

Although there were some significant item-level p-value differences across modes, the magnitude of the p-value differences in Comparability 2 in general was not very large (i.e., < ± .05 for the majority of items, and < ± .10 for almost all items). The largest p-value differences corresponded to absolute effect sizes between 0.25 and 0.30. Most p-value differences corresponded to absolute effect sizes less than 0.20. By Cohen's (1988) standard, an effect size of 0.20 would be considered small, while an effect size of 0.50 would be considered medium.

In all, the findings suggest that for the test forms studied, the observed performance differences might have a fairly small effect in practice. Still, it would be wise to develop an understanding of the factors that can influence examinee behavior and to design a computer interface accordingly, to ensure that examinees are responding to test content rather than features inherent in presenting the test on computer. Information learned about how examinees interact with computer interface features through reviews of the type presented in this paper can help practitioners make decisions about how best to present passage-based tests via computer.

Endnotes

1 Because different sets of items were included in the Mathematics test across the two comparability studies, it was not possible to compare results across the two studies. For the sake of completeness, Mathematics is included in the discussion of the study design, but results for Mathematics are not presented in this paper.

2 Further results for Mathematics will not be presented.

3 For examinees matched to their scores on a nationally standardized achievement test, average Reading and Science Reasoning scores on the national test were at least one scale score point lower for Comparability 2 examinees than for Comparability 1 examinees.

4 Computed as the difference in computer and paper means, divided by the pooled standard deviation.

References

Bergstrom, B. A. (1992, April). *Ability measure equivalence of computer adaptive and pencil and paper tests: A research synthesis.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Bridgeman, B., Lennon, M. L., & Jackenthal, A. (2003). Effects of screen size, screen resolution, and display rate on computer-based test performance. *Applied Measurement in Education, 16*, 191–205.

Choi, S. W., & Tinkler, T. (2002, April). *Evaluating comparability of paper-and-pencil and computer-based assessment in a K–12 setting.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.

Dillon, A. (1992). Reading from paper versus screens: A critical review of the empirical literature. *Ergonomics, 35*, 1297–1326.

Duchnicky, R. L., & Kolers, P. A. (1983). Readability of text scrolled on visual display terminals as a function of window size. *Human Factors, 25*, 683–692.

Fisher, D. L., & Tan, K. C. (1989). Visual displays: The highlighting paradox. *Human Factors, 31*, 17–30.

Hetter, R. D., Segall, D. O., & Bloxom, B. M. (1997). Evaluating item calibration medium in computerized adaptive testing. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized Adaptive Testing: From Inquiry to Operation* (Chapter 16, pp. 161–167). Washington, DC: American Psychological Association.

Mazzeo, J., Druesne, B., Raffeld, P. C., Checketts, K. T., & Muhlstein, A. (1991). *Comparability of computer and paper-and-pencil scores for two CLEP general examinations* (College Board Rep. No. 91–5). New York: College Entrance Examination Board.

Muter, P. (1996). Interface design and optimization of reading of continuous text. In H. van Oostendorp & S. de Mul (Eds.), *Cognitive Aspects of Electronic Text Processing* (pp. 161–180). Norwood, NJ: Ablex.

Muter, P., & Maurutto, P. (1991). Reading and skimming from computer screens and books: The paperless office revisited? *Behaviour and Informational Technology, 10*, 257–266.

Parshall, C. G., Spray, J. A., Kalohn, J. C., & Davey, T. (2002). *Practical considerations in computer-based testing*. New York: Springer.

Pommerich, M., & Burden, T. (2000, April). *From simulation to application: Examinees react to computerized testing*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Russell, M. (1999). Testing on computers: A follow-up study comparing performance on computer and on paper. *Education Policy Analysis Archives, 7*(20). Available online at http://epaa.asu.edu/epaa/v7n20

Russell, M., & Haney, W. (1997). Testing writing on computers: An experiment comparing student performance on tests computed via computer and via paper-and-pencil. *Education Policy Analysis Archives, 5*(3). Available online at http://epaa.asu.edu./epaa/v5n3.html

Sawaki, Y. (2001). Comparability of conventional and computerized tests of reading in a second language. *Language Learning & Technology, 5*, 38–59.

Schaeffer, G. A., Reese, C. M., Steffen, M., McKinley, R. L., & Mills, C. N. (1993). *Field test of a computer-based GRE general test* (GRE Board Rep. No. 88-08P). Princeton, NJ: Educational Testing Service.

Schwarz, E., Beldie, I. P., & Pastoor, S. (1983). A comparison of paging and scrolling for changing screen contents by inexperienced users. *Human Factors, 25*, 279–282.

Spray, J. A., Ackerman, T. A., Reckase, M. D., & Carlson, J. E. (1989). Effect of the medium of item presentation on examinee performance and item characteristics. *Journal of Educational Measurement, 26*, 261–271.

Tullis, T. S. (1997). Screen design. In M. Helander, T. K. Landauer, & P. V. Prabhu (Eds.), *Handbook of Human-Computer Interaction* (2nd ed., pp. 503–531). Amsterdam: North-Holland.

Wang, T., & Kolen, M. J. (2001). Evaluating comparability in computerized adaptive testing: Issues, criteria and an example. *Journal of Educational Measurement, 38*, 19–49.

Author's Note

This paper represents the culmination of several years of research, and incorporates findings across several different studies that were conducted while the author was at ACT, Inc. The research was a collaboration among countless people who contributed to the design and development of the computer interface and tutorials used in the studies, to the planning, design, and data collection for the studies, and to data analyses that were conducted across the studies. Contributors to the project included members of the Support, Technological Applications and Research Department, the Elementary and Secondary School Programs Department, the Measurement Research Department, the Educational Technology Center, the Systems Support Department, the Placement Programs Department, the Operations Department, the Educational Services Department, and the Statistical Research Department, all at ACT, Inc.

The views expressed are those of the author and not necessarily those of the Department of Defense, the United States Government, or ACT, Inc.

Correspondence concerning this article should be addressed to
Mary Pommerich, Defense Manpower Data Center, DoD Center Monterey Bay,
400 Gigling Rd., Seaside, CA 93955-6771. e-mail: pommermr@osd.pentagon.mil

Author Biography

Mary Pommerich is a psychometrician with the ASVAB (Armed Services Vocational Aptitude Battery) testing program. She is interested in the pursuit of quality measurement.

J·T·L·A

The Journal of Technology, Learning, and Assessment

www.jtla.org

Technology and Assessment Study Collaborative
Caroline A. & Peter S. Lynch School of Education, Boston College